Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Confidence Sets for the Optimal Approximating Model – Bridging a Gap between Adaptive Point Estimation and Confidence Regions

Angelika Rohde[1], Lutz Dümbgen[2]

submitted: 26th August 2008

[1] Weierstraß-Institut Berlin
E-Mail: rohde@wias-berlin.de

[2] Universität Bern

No. 1354
Berlin 2008

2000 *Mathematics Subject Classification.* 62G15, 62G20.

Key words and phrases. Adaptivity, confidence sets, coupling, exponential inequality, model selection, multiscale inference, risk optimality..

This work was supported by the Swiss National Science Foundation..

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

In the setting of high-dimensional linear models with Gaussian noise, we investigate the possibility of confidence statements connected to model selection. Although there exist numerous procedures for adaptive point estimation, the construction of adaptive confidence regions is severely limited (cf. Li, 1989). The present paper sheds new light on this gap. We develop exact and adaptive confidence sets for the best approximating model in terms of risk. Our construction is based on a multiscale procedure and a particular coupling argument. Utilizing exponential inequalities for noncentral χ^2–distributions, we show that the risk and quadratic loss of all models within our confidence region are uniformly bounded by the minimal risk times a factor close to one.

1 Introduction

When dealing with a high dimensional observation vector, the natural question arises whether the data generating process can be approximated by a model of substantially lower dimension. Rather than on the true model, the focus is here on smaller ones which still contain the essential information and allow for interpretation. Typically, the models under consideration are characterized by the non-zero components of some parameter vector. Estimating the true model requires the rather idealistic situation that each component is either sufficiently large or equal to zero: A small perturbation of the parameter vector always results in the biggest model, with what the question about the true model does not seem to be adequate in general. Precisely, the model which is optimal in terms of risk then appears as target of many model selection strategies. Within a specified class of competing models, this paper is concerned with confidence regions for that approximating model which is optimal in terms of risk.

Suppose that we observe a random vector $X_n = (X_{in})_{i=1}^n$ with distribution $\mathcal{N}_n(\theta_n, \sigma^2 I_n)$ together with an estimator $\hat{\sigma}_n$ for the standard deviation $\sigma > 0$. Often the signal θ_n represents coefficients of an unknown smooth function with respect to a given orthonormal basis of functions.

There is a vast amount of literature on point estimation of θ_n. For a given estimator $\hat{\theta}_n = \hat{\theta}_n(X_n, \hat{\sigma}_n)$ for θ_n, let

$$L(\hat{\theta}_n, \theta_n) \;:=\; \|\hat{\theta}_n - \theta_n\|_n^2 \quad \text{and} \quad R(\hat{\theta}_n, \theta_n) \;:=\; \mathbb{E}L(\hat{\theta}_n, \theta_n)$$

be its quadratic loss and the corresponding risk, respectively. Here $\|\cdot\|_n$ denotes the standard Euclidean norm of vectors divided by $\sqrt{n}$. Various adaptivity results are known

for this setting, often in terms of oracle inequalities. A typical result reads as follows: Let $(\check{\theta}_n^{(c)})_{c\in\mathcal{C}_n}$ be a family of candidate estimators $\check{\theta}_n^{(c)} = \check{\theta}_n^{(c)}(X_n)$ for θ_n, where $\sigma > 0$ is temporarily assumed to be known. Then there exist estimators $\hat{\theta}_n$ and constants $A_n, B_n = O(\log(n)^\gamma)$ with $\gamma \geq 0$ such that for arbitrary θ_n in a certain set $\Theta_n \subset \mathbb{R}^n$,

$$R(\hat{\theta}_n, \theta_n) \;\leq\; A_n \inf_{c\in\mathcal{C}_n} R(\check{\theta}_n^{(c)}, \theta_n) + \frac{B_n}{n}\sigma^2.$$

Results of this type are provided, for instance, by Polyak and Tsybakov (1991) and Donoho and Johnstone (1994, 1995, 1998). Further results of this type, partly in different settings, have been provided by Stone (1984), Lepski et al. (1997), Efromovich (1998), Cai (1999, 2002), to mention just a few.

By way of contrast, when aiming at adaptive confidence sets one faces severe limitations. Here is a result of Li (1989), slightly rephrased: Suppose that Θ_n contains a closed ball $B(\theta_n^o, cn^{-1/4})$ with respect to $\|\cdot\|_n$, where $c > 0$. Still assuming σ to be known, let $\hat{D}_n = \hat{D}_n(X_n) \subset \Theta_n$ be a $(1-\alpha)$–confidence set for $\theta_n \in \Theta_n$. Such a confidence set may may be used as a test of the (Bayesian) null hypothesis that θ_n is uniformly distributed on the sphere $\partial B(\theta_n^o, cn^{-1/4})$ versus the alternative that $\theta_n = \theta_n^o$: We reject this null hypothesis at level α if $\|\eta - \theta_n^o\|_n < cn^{-1/4}$ for all $\eta \in \hat{D}_n$. Since this test cannot have larger power than the corresponding Neyman-Pearson test,

$$\begin{aligned}\mathbb{P}_{\theta_n^o}\Big(\sup_{\eta\in\hat{D}_n} \|\eta - \theta_n^o\|_n < cn^{-1/4}\Big) \;&\leq\; \mathbb{P}\Big(S_n^2 \leq \chi^2_{n;\alpha}(n^{1/2}c^2/\sigma^2)\Big) \quad (\text{with } S_n^2 \sim \chi_n^2)\\ &=\; \Phi\Big(\Phi^{-1}(\alpha) + 2^{-1/2}c^2/\sigma^2\Big) + o(1),\end{aligned}$$

where $\chi^2_{n;\alpha}(\delta^2)$ stands for the α–quantile of the noncentral chi–squared distribution with n degrees of freedom and noncentrality parameter δ^2. Throughout this paper, asymptotic statements refer to $n \to \infty$. The previous inequality entails that no reasonable confidence set has a diameter of order $o_p(n^{-1/4})$ uniformly over the parameter space Θ_n, as long as the latter is sufficiently large. Despite these limitations, there is some literature on confidence sets in the present or similar settings; see for instance Beran (1996, 2000), Beran and Dümbgen (1998) and Genovese and Wassermann (2005).

Improving the rate of $O_p(n^{-1/4})$ is only possible via additional constraints on θ_n, i.e. considering substantially smaller sets Θ_n. For instance, Baraud (2004) developed nonasymptotic confidence regions which perform well on finitely many linear subspaces. Robins and van der Vaart (2006) construct confidence balls via sample splitting which adapt to some extent to the unknown "smoothness" of θ_n. In their context, Θ_n corresponds to a Sobolev smoothness class with given parameter (β, L). However, adaptation in this context is possible only within a range $[\beta, 2\beta]$. Independently, Cai and Low (2006) treat the same problem in the special case of the Gaussian white noise model, obtaining the same kind of adaptivity in the broader scale of Besov bodies. Other possible constraints on θ_n are so-called shape constraints; see for instance Cai and Low (2007), Dümbgen (2003) or Hengartner and Stark (1995).

The question is whether one can bridge this gap between confidence sets and point estimators. More precisely, we would like to understand the possibility of adaptation for point estimators in terms of some confidence region for the set of all optimal candidate estimators $\check{\theta}_n^{(c)}$. That means, we want to construct a confidence region $\hat{\mathcal{K}}_{n,\alpha} = \hat{\mathcal{K}}_{n,\alpha}(X_n, \hat{\sigma}_n) \subset \mathcal{C}_n$ for the set

$$\mathcal{K}_n(\theta_n) \;:=\; \operatorname*{argmin}_{c \in \mathcal{C}_n} R(\check{\theta}_n^{(c)}, \theta_n)$$

such that for arbitrary $\theta_n \in \mathbb{R}^n$,

$$\mathbb{P}_{\theta_n}\Big(\mathcal{K}_n(\theta_n) \subset \hat{\mathcal{K}}_{n,\alpha}\Big) \;\geq\; 1 - \alpha \tag{1}$$

and

$$\left.\begin{array}{l} \max\limits_{c \in \hat{\mathcal{K}}_{n,\alpha}} R(\check{\theta}_n^{(c)}, \theta_n) \\ \max\limits_{c \in \hat{\mathcal{K}}_{n,\alpha}} L(\check{\theta}_n^{(c)}, \theta_n) \end{array}\right\} \;=\; O_p(A_n) \min_{c \in \mathcal{C}_n} R(\check{\theta}_n^{(c)}, \theta_n) + \frac{O_p(B_n)}{n}\sigma^2. \tag{2}$$

Solving this problem means that statistical inference about differences in the performance of estimators is possible, although inference about their risk and loss is severely limited. In some settings, selecting estimators out of a class of competing estimators entails estimating implicitly an unknown regularity or smoothness class for the underlying signal θ_n. Computing a confidence region for good estimators is particularly suitable in situations in which several good candidate estimators fit the data equally well although they look different. This aspect of exploring various candidate estimators is not covered by the usual theory of point estimation.

Note that our confidence region $\hat{\mathcal{K}}_{n,\alpha}$ is required to contain the whole set $\mathcal{K}_n(\theta_n)$, not just one element of it, with probability at least $1 - \alpha$. The same requirement is used by Futschik (1999) for inference about the argmax of a regression function.

The remainder of this paper is organized as follows. For the reader's convenience our approach is first described in a simple toy model in Section 2. In Section 3 we develop and analyze an explicit confidence region $\hat{\mathcal{K}}_{n,\alpha}$ related to $\mathcal{C}_n := \{0, 1, \ldots, n\}$ with candidate estimators

$$\check{\theta}_n^{(k)} \;:=\; \big(1\{i \leq k\} X_{in}\big)_{i=1}^n.$$

These correspond to a standard nested sequence of approximating models. Section 4 discusses richer families of candidate estimators. All proofs and auxiliary results are deferred to Sections 5 and 6.

2 A toy problem

Suppose we observe a stochastic process $Y = (Y(t))_{t \in [0,1]}$, where

$$Y(t) \;=\; F(t) + W(t), \quad t \in [0, 1],$$

with an unknown fixed continuous function F on $[0,1]$ and a Brownian motion $W = (W(t))_{t\in[0,1]}$. We are interested in the set

$$\mathcal{S}(F) \;:=\; \operatorname*{argmin}_{t\in[0,1]} F(t).$$

Precisely, we want to construct a $(1-\alpha)$–confidence region $\hat{\mathcal{S}}_\alpha = \hat{\mathcal{S}}_\alpha(Y) \subset [0,1]$ for $\mathcal{S}(F)$ in the sense that

$$P\big(\mathcal{S}(F) \subset \hat{\mathcal{S}}_\alpha\big) \;\geq\; 1-\alpha, \tag{3}$$

regardless of F. To construct such a confidence set we regard $Y(s) - Y(t)$ for arbitrary different $s,t \in [0,1]$ as a test statistic for the null hypothesis that $s \in \mathcal{S}$, i.e. large values of $Y(s) - Y(t)$ give evidence for $s \notin \mathcal{S}$.

A first naive proposal is the set

$$\hat{\mathcal{S}}_\alpha^{\text{naive}} \;:=\; \Big\{ s \in [0,1] : Y(s) \leq \min_{[0,1]} Y + \kappa_\alpha^{\text{naive}} \Big\}$$

with $\kappa_\alpha^{\text{naive}}$ denoting the $(1-\alpha)$–quantile of $\max_{[0,1]} W - \min_{[0,1]} W$.

Here is a refined version based on results of Dümbgen and Spokoiny (2001): Let κ_α be the $(1-\alpha)$–quantile of

$$\sup_{s,t\in[0,1]} \left(\frac{|W(s) - W(t)|}{\sqrt{|s-t|}} - \Gamma(s-t) \right), \tag{4}$$

where

$$\Gamma(\pm\delta) \;:=\; \sqrt{2\log(e/\delta)} \quad \text{for } 0 \leq \delta \leq 1.$$

Then constraint (3) is satisfied by the confidence region

$$\hat{\mathcal{S}}_\alpha \;:=\; \Big\{ s \in [0,1] : Y(s) \leq Y(t) + \sqrt{|s-t|}\big(\Gamma(s-t) + \kappa_\alpha\big) \text{ for all } t \in [0,1] \Big\}.$$

To illustrate the power of this method, consider for instance a sequence of functions $F = F_n$ such that for some parameters $s_n \in [0,1]$, $\gamma > 1/2$ and $c_n \to \infty$,

$$F_n(t) - F_n(s_n) \;\geq\; c_n |t - s_n|^\gamma \quad \text{for all } t \in [0,1].$$

Then for the naive confidence region one can only deduce that

$$\max_{t\in\hat{\mathcal{S}}_\alpha^{\text{naive}}} |t - s_n| \;=\; O_p\big(c_n^{-1/\gamma}\big),$$

whereas

$$\max_{t\in\hat{\mathcal{S}}_\alpha} |t - s_n| \;=\; O_p\Big(\log(c_n)^{1/(2\gamma-1)} c_n^{-1/(\gamma-1/2)}\Big).$$

3 Confidence regions for nested approximating models

As in the introduction let $X_n = \theta_n + \epsilon_n$ denote the n-dimensional observation vector with $\theta_n \in \mathbb{R}^n$ and $\epsilon_n \sim \mathcal{N}_n(0, \sigma^2 I_n)$. For any candidate estimator $\check{\theta}_n^{(k)} = \big(1\{i \le k\}X_{in}\big)_{i=1}^n$ the loss is given by

$$L_n(k) := L(\check{\theta}_n^{(k)}, \theta_n) \;=\; \frac{1}{n}\sum_{i=k+1}^{n} \theta_{in}^2 + \frac{1}{n}\sum_{i=1}^{k}(X_{in} - \theta_{in})^2$$

with corresponding risk

$$R_n(k) := R(\check{\theta}_n^{(k)}, \theta_n) \;=\; \frac{1}{n}\sum_{i=k+1}^{n} \theta_{in}^2 + \frac{k}{n}\sigma^2.$$

Model selection usually aims at estimating a candidate estimator which is optimal in terms of risk. Since the risk depends on the unknown signal and therefore is not available, the selection procedure minimizes an unbiased risk estimator instead. In the sequel, the bias-corrected risk estimator for the candidate $\check{\theta}_n^{(k)}$ is defined as

$$\hat{R}_n(k) \;:=\; \frac{1}{n}\sum_{i=k+1}^{n} (X_{in}^2 - \hat{\sigma}_n^2) + \frac{k}{n}\hat{\sigma}_n^2,$$

where $\hat{\sigma}_n^2$ is a variance estimator satisfying the subsequent condition.

(A) $\hat{\sigma}_n^2$ and X_n are stochastically independent with

$$\frac{m\hat{\sigma}_n^2}{\sigma^2} \;\sim\; \chi_m^2,$$

where $1 \le m = m_n \le \infty$ with $m = \infty$ meaning that σ is known, i.e. $\hat{\sigma}_n^2 \equiv \sigma^2$. For asymptotic statements, it is generally assumed that

$$\beta_n^2 \;:=\; \frac{2n}{m_n} \;=\; O(1)$$

unless stated otherwise.

Example Suppose that we observe $Y = M\eta + \delta$ with given design matrix $M \in \mathbb{R}^{(n+m)\times n}$ of rank n, unknown parameter vector $\eta \in \mathbb{R}^n$ and unobserved error vector $\delta \sim \mathcal{N}_{n+m}(0, \sigma^2 I_{n+m})$. Then the previous assumptions are satisfied by $X_n := (M^\top M)^{1/2}\hat{\eta}$ with $\hat{\eta} := (M^\top M)^{-1}M^\top Y$ and $\hat{\sigma}_n^2 := \|Y - M\hat{\eta}\|^2/m$, where $\theta_n := (M^\top M)^{1/2}\eta$.

Important for our analysis is the behavior of the difference process

$$D_n = (D_n(j,k))_{0\le j<k\le n} \;:=\; \frac{\sqrt{n}}{\hat{\sigma}_n^2}\Big(\big(\hat{R}_n(k) - R_n(k)\big) - \big(\hat{R}_n(j) - R_n(j)\big)\Big)_{0\le j<k\le n}.$$

Its asymptotic distribution depends on the unknown "signal-to-noise" vector $(\theta_{in}^2/\sigma^2)_{i=1}^n$, as seen in the following proposition. It provides an approximation of the difference process by a Gaussian process, where approximation in distribution refers to the dual bounded Lipschitz metric d_w, which metrizes the weak topology. For details we refer to Section 6.

Proposition 1. *In case of $\|\theta_n\|_n^2 = O(1)$, the difference process D_n is approximated in distribution by a centered Gaussian process Δ_n with covariances*

$$\operatorname{cov}\Big(\Delta_n(j,k), \Delta_n(j',k')\Big) = \frac{4\beta_n^2}{n^2}(k-j)(k'-j') + \frac{1}{n}\sum_{i\in(j,k]\cap(j',k']}\left(4\frac{\theta_{in}^2}{\sigma^2}+2\right).$$

If we could estimate the covariance function in Proposition 1 consistently, we could imitate the naive confidence region of Section 2. For a more powerful confidence region, the crucial step is to analyze a suitably standardized version of the increment process D_n, getting additionally rid of the restriction on $\|\theta_n\|_n^2$. Since this process does not have subgaussian tails, the standardization is more involved than the correction in (4).

Theorem 2. *Let $(\theta_n)_{n\in\mathbb{N}}$ be arbitrary. For $0 \le j < k \le n$ let*

$$\gamma_n(j,k)^2 \;:=\; \frac{1}{n}\sum_{i=j+1}^{k}(4\,\theta_{in}^2/\sigma^2+2)$$

and define

$$C_{jkn} \;:=\; \left(1+\frac{5\,\Gamma_{jkn}}{\sqrt{n}\,\gamma_n(j,k)}\right)\Gamma_{jkn}$$

with $\Gamma_{jkn} := \Gamma\big(\max\big\{\gamma_n(j,k)^2/\gamma_n(0,n)^2,\, 1/n\big\}\big)$. Then the sequence of random variables

$$d_n \;:=\; \sup_{0\le j<k\le n}\left(\frac{|D_n(j,k)|}{\gamma_n(j,k)} - C_{jkn}\right)$$

is tight. Precisely, it is approximated in distribution by

$$\delta_n \;:=\; \sup_{0\le j<k\le n}\left(\frac{\big|\,W(\gamma_n(0,k)^2) - W(\gamma_n(0,j)^2) + 2\beta_n(k-j)/n\,Z\,\big|}{\gamma_n(j,k)} - \Gamma_{jkn}\right),$$

where W denotes a Brownian motion, independent of $Z\sim\mathcal{N}(0,1)$.

The above non-degenerate limiting distribution demonstrates that the additive correction is appropriately defined and cannot be chosen essentially smaller.

In order to construct a confidence set for $\mathcal{K}_n(\theta_n)$ by means of d_n, we are facing the problem that the auxiliary function γ_n depends on the unknown signal-to-noise vector θ_n/σ. In fact, knowing γ_n would imply knowledge of $\mathcal{K}_n(\theta_n)$ already. A natural approach is to replace the quantities which are dependent on the unknown parameter by suitable estimates. A common estimator of the variance γ_n^2 is given by

$$\hat\gamma_n(j,k)^2 := \frac{1}{n}\sum_{i=j+1}^{k}\big(4(X_{in}^2/\hat\sigma_n^2-1)+2\big), \quad j<k.$$

However, using such an estimator does not seem to work since

$$\sup_{0\le j<k\le n} \Big| \frac{\hat\gamma_n(j,k)}{\gamma_n(j,k)} - 1 \Big| \;\not\longrightarrow_p\; 0$$

as n goes to infinity. This can be verified by noting that $\big(\hat\gamma_n(j,k)^2\big)_{0\le j<k\le n}$ is, up to centering, essentially of the same structure as the difference process D_n itself.

The least favourable case of constant risk

The problem of estimating the set $\arg\min_k R_n(k)$ can be cast into our toy model where $Y(t)$, $F(t)$ and $W(t)$ correspond to $\hat R_n(k)$, $R_n(k)$ and the difference $\hat R_n(k) - R_n(k)$, respectively. One may expect that the more distinctive the global minima are, the easier it is to identify their location. Hence the case of constant risks appears to be least favourable, corresponding to a signal

$$\theta_n^* \;:=\; \big(\pm\sigma\big)_{i=1}^n,$$

In this situation, each candidate estimator $\check\theta_n^{(k)}$ has the same risk σ^2.

A related consideration leading to an explicit procedure is as follows: For fixed indices $0 \le j < k \le n$,

$$R_n(j) - R_n(k) \;=\; \frac{1}{n}\sum_{i=j+1}^{k} \theta_{in}^2 - \frac{k-j}{n}\sigma^2,$$

and if assumption (A) is satisfied, the statistic

$$T_{jkn} \;:=\; \frac{\sum_{i=j+1}^k X_{in}^2}{(k-j)\hat\sigma_n^2} \;=\; 2 - \frac{n(\hat R_n(k) - \hat R_n(j))}{(k-j)\hat\sigma_n^2}$$

has a noncentral (in the numerator) F–distribution

$$F_{k-j,m}\bigg(\frac{\sum_{i=j+1}^k \theta_{in}^2}{\sigma^2}\bigg) = F_{k-j,m}\bigg(k-j+\frac{n\big(R_n(j)-R_n(k)\big)}{\sigma^2}\bigg)$$

with $k-j$ and m degrees of freedom. Thus large or small values of T_{jkn} give evidence for $R_n(j)$ being larger or smaller, respectively, than $R_n(k)$. Precisely,

$$\mathcal{L}_{\theta_n}(T_{jkn}) \begin{cases} \le_{\text{st.}}\ \mathcal{L}_{\theta_n^*}(T_{jkn}) & \text{whenever } j \in \mathcal{K}_n(\theta_n), \\ \ge_{\text{st.}}\ \mathcal{L}_{\theta_n^*}(T_{jkn}) & \text{whenever } k \in \mathcal{K}_n(\theta_n). \end{cases}$$

Note that this stochastic ordering remains valid if $\hat\sigma_n^2$ is just independent from X_n, i.e. also under the more general requirement of the remark at the end of this section. Via suitable coupling of Poisson mixtures of central χ^2–distributed random variables, this observation is extended to

Proposition 3 (Coupling). *For any $\theta_n \in \mathbb{R}^n$ there exists a probability space with random variables $\big(\tilde{T}_{jkn}\big)_{0\le j<k\le n}$ and $\big(\tilde{T}^*_{jkn}\big)_{0\le j<k\le n}$ such that*

$$\begin{aligned}\mathcal{L}\Big(\big(\tilde{T}_{jkn}\big)_{0\le j<k\le n}\Big) &= \mathcal{L}_{\theta_n}\Big(\big(T_{jkn}\big)_{0\le j<k\le n}\Big),\\ \mathcal{L}\Big(\big(\tilde{T}^*_{jkn}\big)_{0\le j<k\le n}\Big) &= \mathcal{L}_{\theta^*_n}\Big(\big(T_{jkn}\big)_{0\le j<k\le n}\Big),\end{aligned}$$

and for arbitrary indices $0 \le j < k \le n$,

$$\tilde{T}_{jkn} \begin{cases} \le \tilde{T}^*_{jkn} & \text{whenever } j \in \mathcal{K}_n(\theta_n),\\ \ge \tilde{T}^*_{jkn} & \text{whenever } k \in \mathcal{K}_n(\theta_n).\end{cases}$$

As a consequence of Proposition 3, we can define a confidence set for $\mathcal{K}_n(\theta_n)$, based on this least favourable case. Let $\kappa_{n,\alpha}$ denote the $(1-\alpha)$-quantile of $\mathcal{L}_{\theta^*_n}(d_n)$. Motivated by the procedure in Section 2 and Theorem 2, we define

$$\begin{aligned}\hat{\mathcal{K}}_{n,\alpha} &:= \Big\{j:\ \hat{R}_n(j) \le \hat{R}_n(k) + \frac{\hat{\sigma}_n^2\sqrt{6|k-j|}}{n}\Big(\Gamma\Big(\frac{k-j}{n}\Big) + \kappa_{n,\alpha}\Big) \\ &\qquad + \frac{5\hat{\sigma}_n^2}{n}\Gamma\Big(\frac{k-j}{n}\Big)^2 \quad \text{for all } k \ne j\Big\} \qquad (5)\\ &= \Big\{j:\ T_{ijn} \ge 2 - c_{ijn} \text{ for } 1 \le i < j,\ T_{jkn} \le 2 + c_{jkn} \text{ for } j < k \le n\Big\},\end{aligned}$$

with

$$c_{jkn} := \sqrt{\frac{6}{|k-j|}}\Big(\Gamma\Big(\frac{k-j}{n}\Big) + \kappa_{n,\alpha}\Big) + \frac{5}{|k-j|}\Gamma\Big(\frac{k-j}{n}\Big)^2.$$

Theorem 4. *Let $(\theta_n)_{n\in\mathbb{N}}$ be arbitrary. With $\hat{\mathcal{K}}_{n,\alpha}$ as defined above,*

$$\mathbb{P}_{\theta_n}\Big(\mathcal{K}_n(\theta_n) \not\subset \hat{\mathcal{K}}_{n,\alpha}\Big) \le \alpha.$$

In case of $n/m \to 0$, the critical values $\kappa_{n,\alpha}$ converge to the critical value κ_α introduced in Section 2. Under the weaker assumption that $n/m = O(1)$, $\kappa_{n,\alpha} = O(1)$, and the confidence regions $\hat{\mathcal{K}}_{n,\alpha}$ satisfy the oracle inequalities

$$\max_{k\in\hat{\mathcal{K}}_{n,\alpha}} R_n(k) \le \min_{j\in\mathcal{C}_n} R_n(j) + \Big(2\sqrt{15} + o_p(1)\Big)\sqrt{\nu_n \min_{j\in\mathcal{C}_n} R_n(j)} + O_p(\nu_n) \qquad (6)$$

and

$$\max_{k\in\hat{\mathcal{K}}_{n,\alpha}} L_n(k) \le \min_{j\in\mathcal{C}_n} L_n(j) + O_p\Big(\sqrt{\nu_n \min_{j\in\mathcal{C}_n} L_n(j)}\Big) + O_p(\nu_n) \qquad (7)$$

with $\nu_n = (\sigma^2 \log n)/n$ and C some universal constant independent of σ^2.

REMARK (Variance estimation) Instead of condition (A), one may require more generally that $\hat{\sigma}_n^2$ and X_n are independent with

$$\sqrt{n}\Big(\frac{\hat{\sigma}_n^2}{\sigma^2} - 1\Big) \;\to_D\; \mathcal{N}(0, \beta^2)$$

for a given $\beta \geq 0$. This covers, for instance, estimators used in connection with wavelets. There σ is estimated by the median of some high frequency wavelet coefficients divided by the normal quantile $\Phi^{-1}(3/4)$. Theorem 2 continues to hold, and the coupling extends to this situation, too, with S^2 in the proof being distributed as $n\hat{\sigma}_n^2$. Under this assumption on the external variance estimator, the confidence region $\hat{\mathcal{K}}_{n,\alpha}$, defined with $m := \lfloor 2n/\beta^2 \rfloor$, is at least asymptotically valid and satisfies the above oracle inequalities as well.

4 Confidence sets in case of larger families of candidates

The previous result relies strongly on the assumption of nested models. It is possible to obtain confidence sets for the optimal approximating models in a more general setting, albeit the resulting oracle property is not as strong as in the nested case. In particular, we can no longer rely on a coupling result but need a different construction. For the reader's convenience, we focus on the case of known σ, i.e. $m = \infty$; see also the remark at the end of this section.

Let $\mathcal{C}_n$ be a family of index sets $C \subset \{1, 2, \ldots, n\}$ with candidate estimators

$$\check{\theta}^{(C)} \;:=\; \big(1\{i \in C\}X_{in}\big)_{i=1}^n$$

and corresponding risks

$$R_n(C) := R(\check{\theta}^{(C)}, \theta_n) \;=\; \frac{1}{n}\sum_{i \notin C} \theta_{in}^2 + \frac{\#C}{n}\sigma^2.$$

For two index sets C and D,

$$(n/\sigma^2)\big(R_n(D) - R_n(C)\big) \;=\; \delta_n^2(C \setminus D) - \delta_n^2(D \setminus C) + \#D - \#C$$

with the auxiliary quantities

$$\delta_n^2(J) \;:=\; \sum_{i \in J} \theta_{in}^2/\sigma^2, \quad J \subset \{1, 2, \ldots, n\}.$$

Hence we aim at simultaneous $(1-\alpha)$–confidence intervals for these noncentrality parameters $\delta_n(J)$, where $J \in \mathcal{M}_n := \{D \setminus C : C, D \in \mathcal{C}_n\}$. To this end we utilize the fact that

$$T_n(J) \;:=\; \frac{1}{\sigma^2}\sum_{i \in J} X_{in}^2$$

has a $\chi^2_{\#J}(\delta_n^2(J))$-distribution. We denote the distribution function of $\chi^2_k(\delta^2)$ by $F_k(\cdot \mid \delta^2)$. Now let $M_n := \#\mathcal{M}_n - 1 \le \#\mathcal{C}_n(\#\mathcal{C}_n - 1)$, the number of nonvoid index sets $J \in \mathcal{M}_n$. Then with probability at least $1-\alpha$,

$$\alpha/(2M_n) \;\le\; F_{\#J}\big(T_n(J) \,\big|\, \delta_n^2(J)\big) \;\le\; 1-\alpha/(2M_n) \quad \text{for all } J \in \mathcal{M}_n, J \neq \emptyset. \tag{8}$$

Since $F_{\#J}(T_n(J) \mid \delta^2)$ is strictly decreasing in δ^2 with limit 0 as $\delta^2 \to \infty$, (8) entails the following simultaneous $(1-\alpha)$–confidence intervals $\big[\hat\delta^2_{n,\alpha,l}(J), \hat\delta^2_{n,\alpha,u}(J)\big]$ for all parameters $\delta_n^2(J)$: We set $\hat\delta^2_{n,\alpha,l}(\emptyset) := \hat\delta^2_{n,\alpha,u}(\emptyset) := 0$, while for nonvoid J,

$$\hat\delta^2_{n,\alpha,l}(J) \;:=\; \min\Big\{\delta^2 \ge 0 : F_{\#J}\big(T_n(J) \,\big|\, \delta^2\big) \le 1-\alpha/(2M_n)\Big\}, \tag{9}$$

$$\hat\delta^2_{n,\alpha,u}(J) \;:=\; \max\Big\{\delta^2 \ge 0 : F_{\#J}\big(T_n(J) \,\big|\, \delta^2\big) \ge \alpha/(2M_n)\Big\}. \tag{10}$$

By means of these bounds, we may claim with confidence $1-\alpha$ that for arbitrary $C, D \in \mathcal{C}_n$ the normalized difference $(n/\sigma^2)\big(R_n(D) - R_n(C)\big)$ is at most $\hat\delta^2_{n,\alpha,u}(C\setminus D) - \hat\delta^2_{n,\alpha,l}(D\setminus C) + \#D - \#C$. Thus a $(1-\alpha)$–confidence set for $\mathcal{K}_n(\theta_n) = \operatorname{argmin}_{C\in\mathcal{C}_n} R_n(C)$ is given by

$$\hat{\mathcal{K}}_{n,\alpha} \;:=\; \Big\{C \in \mathcal{C}_n : \hat\delta^2_{n,\alpha,u}(C \setminus D) - \hat\delta^2_{n,\alpha,l}(D \setminus C) + \#D - \#C \ge 0 \text{ for all } D \in \mathcal{C}_n\Big\}.$$

These confidence sets $\hat{\mathcal{K}}_{n,\alpha}$ satisfy the following oracle inequalities:

Theorem 5. *Let $(\theta_n)_{n\in\mathbb{N}}$ be arbitrary, and suppose that $\log \#\mathcal{C}_n = o(n)$. Then*

$$\max_{C\in\hat{\mathcal{K}}_{n,\alpha}} R_n(C) \;\le\; \min_{D\in\mathcal{C}_n} R_n(D) \;+\; O_p\Big(\sqrt{\tilde\nu_n \min_{D\in\mathcal{C}_n} R_n(D)}\Big) \;+\; O_p\big(\tilde\nu_n\big),$$

$$\max_{C\in\hat{\mathcal{K}}_{n,\alpha}} L_n(C) \;\le\; \min_{D\in\mathcal{C}_n} L_n(D) \;+\; O_p\Big(\sqrt{\tilde\nu_n \min_{D\in\mathcal{C}_n} L_n(D)}\Big) \;+\; O_p\big(\tilde\nu_n\big)$$

with $\tilde\nu_n := \sigma^2 \log(\#\mathcal{C}_n)/n$.

REMARK The upper bounds in Theorem 5 are of the form $\rho_n + O_p\big(\sqrt{\rho_n\tilde\nu_n}\big) + O_p\big(\tilde\nu_n\big)$, with ρ_n denoting minimal risk or minimal loss. For any fixed $\varepsilon > 0$ this bound doesn't exceed $(1+\varepsilon)\rho_n + O_p(\tilde\nu_n)$. Thus Theorem 5 entails that the maximal risk (loss) over $\hat{\mathcal{K}}_{n,\alpha}$ exceed the minimal risk (loss) by a factor close to one, provided that the minimal risk (loss) is substantially larger than $\tilde\nu_n$.

REMARK (Suboptimality in case of nested models) In case of nested models, the general construction is suboptimal in the factor of the leading (in most cases) term $\sqrt{\min_j R_n(j)}$; following the proof carefully and using $\tilde\nu_n = 2\nu_n + O(1)$ in this special setting, one may verify that

$$\max_{k\in\hat{\mathcal{K}}_{n,\alpha}} R_n(k) \;\le\; \min_{j\in\mathcal{C}_n} R_n(j) \;+\; \big(8\sqrt{2} + o_p(1)\big)\sqrt{\nu_n \min_{j\in\mathcal{C}_n} R_n(j)} \;+\; O_p\big(\nu_n\big).$$

The intrinsic reason seems to be that the general procedure does not assume any structure of the candidate estimators so that advanced multiscale theory is not applicable.

REMARK In case of unknown σ, let $\alpha' := 1-(1-\alpha)^{1/2}$. Then with probability at least $1-\alpha'$,

$$\alpha'/2 \;\le\; F_m\big(m(\hat\sigma_n/\sigma)^2 \,\big|\, 0\big) \;\le\; 1-\alpha'/2.$$

The latter inequalities entail that $(\sigma/\hat\sigma_n)^2$ lies between $\tau_{n,\alpha,l} := m/\chi_{m;1-\alpha'/2}$ and $\tau_{n,\alpha,u} := m/\chi^2_{m;\alpha'/2}$. Then we obtain simultaneous $(1-\alpha)$–confidence bounds $\hat\delta^2_{n,\alpha,l}(J)$ and $\hat\delta^2_{n,\alpha,u}(J)$ as in (9) and (10) by replacing α with α' and $T_n(J)$ with

$$\frac{\tau_{n,\alpha,l}}{\hat\sigma_n^2}\sum_{i\in J} X_{in}^2 \quad\text{and}\quad \frac{\tau_{n,\alpha,u}}{\hat\sigma_n^2}\sum_{i\in J} X_{in}^2,$$

respectively. The conclusions of Theorem 5 continue to hold, as long as $n/m_n = O(1)$.

5 Proofs

5.1 Exponential inequalities

An essential ingredient for our main results is an exponential inequality for quadratic functions of a Gaussian random vector. It extends inequalities of Dahlhaus and Polonik (2006) for quadratic forms and may be of independent interest.

Proposition 6. *Let $Z_1, Z_2, \ldots, Z_n$ be independent standard normally distributed random variables. Furthermore, let $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\delta_1, \delta_2, \ldots, \delta_n$ be real constants, and define $\gamma^2 := \mathrm{Var}\Big(\sum_{i=1}^n \lambda_i (Z_i+\delta_i)^2\Big) = \sum_{i=1}^n \lambda_i^2(2+4\delta_i^2)$. Then*

$$\begin{aligned}\mathbb{P}\Big(\sum_{i=1}^n \lambda_i\big((Z_i+\delta_i)^2 - (1+\delta_i^2)\big) \ge \eta\gamma\Big) &\le \exp\Big(-\frac{\eta^2/2}{1+2\eta\lambda_{\max}/\gamma}\Big)\\ &\le e^{1/8}\exp(-\eta/4)\end{aligned}$$

for arbitrary $\eta \ge 0$, where $\lambda_{\max} := \max(\lambda_1, \lambda_2, \ldots, \lambda_n, 0)$.

Note that replacing λ_i in Proposition 6 with $-\lambda_i$ yields twosided exponential inequalities. By means of Proposition 6 and elementary calculations one obtains exponential and related inequalities for noncentral χ^2 distributions:

Corollary 7. *For an integer $n > 0$ and a constant $\delta \ge 0$ let $F_n(\cdot \mid \delta^2)$ be the distribution function of $\chi^2_n(\delta^2)$. Then for arbitrary $r \ge 0$,*

$$F_n(n+\delta^2+r \mid \delta^2) \;\ge\; 1-\exp\Big(-\frac{r^2}{4n+8\delta^2+4r}\Big), \tag{11}$$

$$F_n(n+\delta^2-r \mid \delta^2) \;\le\; \exp\Big(-\frac{r^2}{4n+8\delta^2}\Big). \tag{12}$$

In particular, for any $\alpha \in (0,1)$ *and* $A := \log(2/\alpha)$,

$$F_n^{-1}(1-\alpha/2 \mid \delta^2) \;\leq\; n + \delta^2 + \sqrt{(4n+8\delta^2)A} + 4A, \tag{13}$$

$$F_n^{-1}(\alpha/2 \mid \delta^2) \;\geq\; n + \delta^2 - \sqrt{(4n+8\delta^2)A}. \tag{14}$$

Moreover, for any number $\hat{\delta} \geq 0$, *the inequalities* $\alpha/2 \leq F_n(t \mid \delta^2) \leq 1-\alpha/2$ *entail that*

$$\hat{\delta}^2 - \sqrt{(4n+8\hat{\delta}^2)A} \;\leq\; \delta^2 \;\leq\; \hat{\delta}^2 + \sqrt{(4n+8\hat{\delta}^2)A} + 8A. \tag{15}$$

Conclusion (15) follows from (11) and (12), applied to $r = \hat{\delta}^2 - \delta^2$ and $r = \delta^2 - \hat{\delta}^2$, respectively.

PROOF OF PROPOSITION 6 Standard calculations show that for $0 \leq t < (2\lambda_{\max})^{-1}$,

$$\mathbb{E}\exp\Big(t\sum_{i=1}^n \lambda_i (Z_i+\delta_i)^2\Big) \;=\; \exp\Big(\frac{1}{2}\sum_{i=1}^n \Big\{\delta_i^2 \frac{2t\lambda_i}{1-2t\lambda_i} - \log(1-2t\lambda_i)\Big\}\Big).$$

Then for any such t,

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^n \lambda_i\big((Z_i+\delta_i)^2 - (1+\delta_i^2)\big) \geq \eta\gamma\Big) \\
&\leq\; \exp\Big(-t\eta\gamma - t\sum_{i=1}^n \lambda_i(1+\delta_i^2)\Big)\cdot \mathbb{E}\exp\Big(t\sum_{i=1}^n \lambda_i (Z_i+\delta_i)^2\Big) \\
&=\; \exp\Big(-t\eta\gamma + \frac{1}{2}\sum_{i=1}^n \Big\{\delta_i^2 \frac{4t^2\lambda_i^2}{1-2t\lambda_i} - \log(1-2t\lambda_i) - 2t\lambda_i\Big\}\Big).
\end{aligned} \tag{16}$$

Since the derivative of $x \mapsto -\log(1-x) - x$ equals $x/(1-x)$, one can easily deduce that

$$-\log(1-x) - x \;\leq\; \begin{cases} x^2/2 & \text{if } x \leq 0, \\ x^2/(2(1-x)) & \text{if } x \geq 0. \end{cases}$$

Thus (16) is not greater than

$$\exp\Big(-t\eta\gamma + \frac{1}{2}\sum_{i=1}^n \Big\{\delta_i^2 \frac{4t^2\lambda_i^2}{1-2t\lambda_i} + \frac{2t^2\lambda_i^2}{1-2t\max(\lambda_i,0)}\Big\}\Big) \;\leq\; \exp\Big(-t\eta\gamma + \frac{\gamma^2 t^2/2}{1-2t\lambda_{\max}}\Big).$$

Setting

$$t \;:=\; \frac{\eta}{\gamma + 2\eta\lambda_{\max}} \;\in\; \big[0, (2\lambda_{\max})^{-1}\big),$$

the preceding bound becomes

$$\mathbb{P}\Big(\sum_{i=1}^n \lambda_i\big((Z_i+\delta_i)^2 - (1+\delta_i^2)\big) \geq \eta\gamma\Big) \;\leq\; \exp\Big(-\frac{\eta^2/2}{1+2\eta\lambda_{\max}/\gamma}\Big).$$

Finally, since $\lambda_{\max} \leq \gamma$, the second asserted inequality follows from

$$\frac{\eta^2/2}{1+2\eta\lambda_{\max}/\gamma} \;\geq\; \frac{\eta^2/2}{1+2\eta} \;=\; \frac{\eta}{4} - \frac{\eta}{4+8\eta} \;\geq\; \frac{\eta}{4} - \frac{1}{8}. \qquad \square$$

5.2 Proofs of the main results

For notational convenience, we denote by X_k and ϵ_k the k-th component of the n-dimensional observation and error vector respectively and drop the index n if this is clear from the context. Throughout the proofs, let $\mathcal{T}_n := \big\{(j,k) \;\big|\; 0 \le j < k \le n\big\}$.

PROOF OF PROPOSITION 1 Because $\hat{\sigma}_n^2/\sigma^2 \to_p 1$, it is sufficient to prove the result for

$$\tilde{D}_n \;:=\; \frac{\hat{\sigma}_n^2}{\sigma^2} D_n,$$

where we may assume without loss of generality $\sigma^2 = 1$. The process $\tilde{D}_n$, evaluated at some point $(j,k) \in \mathcal{T}_n$, is then given by

$$\begin{aligned}
\tilde{D}_n(j,k) &:= \frac{1}{\sqrt{n}} \sum_{i=j+1}^{k} \big(2\epsilon_i\theta_{in} + (\epsilon_i^2 - \hat{\sigma}_n^2)\big) - \frac{k-j}{\sqrt{n}}\big(\hat{\sigma}_n^2 - 1\big) \\
&= 2\big(1 - \hat{\sigma}_n^2\big)\frac{k-j}{\sqrt{n}} + \frac{1}{\sqrt{n}} \sum_{i=j+1}^{k} \big(2\epsilon_i\theta_{in} + (\epsilon_i^2 - 1)\big) \\
&= \tilde{D}_{1,n}(j,k) + \tilde{D}_{2,n}(j,k).
\end{aligned}$$

Note that both processes are centered with covariance functions

$$\operatorname{cov}\big(\tilde{D}_{1,n}(j,k), \tilde{D}_{1,n}(j',k')\big) = \frac{4}{n^2}\beta_n^2 (k-j)(k'-j')$$

and

$$\operatorname{cov}\big(\tilde{D}_{2,n}(j,k), \tilde{D}_{2,n}(j',k')\big) = \frac{2}{n} \sum_{i \in (j,k] \cap (j',k']} \big(2\theta_{in}^2 + 1\big),$$

respectively. By assumption, the processes $\tilde{D}_{1,n}$ and $\tilde{D}_{2,n}$ are independent. The approximation of the first process is straightforward and hence omitted in this section. In order to investigate

$$\left(\frac{1}{\sqrt{n}} \sum_{i \in (j,k]} \big(2\epsilon_i\theta_{in} + (\epsilon_i^2 - 1)\big)\right)_{(j,k) \in \mathcal{T}_n},$$

we consider in view of Theorem 2 the normed version of the process, i.e. $\tilde{D}_{2,n}/\gamma_n(0,n)$, without any restrictions on $(\theta_n)_{n\in\mathbb{N}}$. In case $\|\theta_n\|_n^2 = O(1)$, $\gamma_n(0,n)$ is uniformly bounded away from zero and infinity, and the result as stated in Proposition 1 follows in particular.

For let ϕ_i be the i'th summand of this process, that is

$$\phi_i(j,k) \;:=\; \frac{I_{(j,k]}(i)}{\sqrt{n}\gamma_n(0,n)}\Big\{2\epsilon_i\theta_{in} + (\epsilon_i^2 - 1)\Big\}.$$

Now define the partition of $\{1,...,n\} =: \mathcal{S}_n = \mathcal{S}_n^{(1)} + \mathcal{S}_n^{(2)}$ with $\mathcal{S}_n^{(1)} := \{i \in \mathcal{S}_n | \theta_{in}^2 \leq \sqrt{n}\gamma_n(0,n)\}$ and $\mathcal{S}_n^{(2)} = \mathcal{S}_n \setminus \mathcal{S}_n^{(1)}$. Then the process $\tilde{D}_{2,n}$ is the sum of the two independent parts

$$\left(\sum_{i\in\mathcal{S}_n^{(1)}} \phi_i(j,k)\right)_{(j,k)\in\mathcal{T}_n} \quad \text{and} \quad \left(\sum_{i\in\mathcal{S}_n^{(2)}} \phi_i(j,k)\right)_{(j,k)\in\mathcal{T}_n}.$$

By Markov's inequality,

$$\mathbb{P}\Big(\sup_{(j,k)\in\mathcal{T}_n} \Big| \sum_{i\in\mathcal{S}_n^{(2)}} \frac{1}{\sqrt{n}\gamma_n(0,n)} I_{(j,k]}(i)(\epsilon_i^2 - \sigma^2) \Big| > \epsilon\Big)$$
$$\leq \frac{1}{\epsilon}\frac{\sharp\mathcal{S}_n^{(2)}}{\sqrt{n}\gamma_n(0,n)}\,\mathbb{E}\Big(\sup_{(j,k)\in\mathcal{T}_n} \Big| \frac{1}{\sharp\mathcal{S}_n^{(2)}} \sum_{i\in\mathcal{S}_n^{(2)}} I_{(j,k]}(i)(\epsilon_i^2 - \sigma^2) \Big|\Big)$$

for any $\varepsilon > 0$. Let $\mathcal{F}_n := \big\{I_{(j,k]} \,\big|\, (j,k) \in \mathcal{T}_n\big\}$. Using Lemma 6.4 in Beran and Dümbgen (1998), the last expression is bounded by

$$\frac{1}{\epsilon}\frac{\sqrt{\sharp\mathcal{S}_n^{(2)}}}{\sqrt{n}\gamma_n(0,n)} C\mathcal{J}(\mathcal{F}_n)\Big\{\mathbb{E}\Big(\frac{1}{\sharp\mathcal{S}_n^{(2)}}\sum_{i\in\mathcal{S}_n^{(2)}}\epsilon_i^4\Big)\Big\}^{1/2},$$

where C denotes a universal constant independent of n and $\mathcal{J}(\mathcal{F}_n)$ stands for $\int_0^1 \sqrt{\log N(u,\mathcal{F}_n)}du$ with $N(u,\mathcal{F}_n)$ the uniform covering number as defined in section 6. Note that for the classes under consideration, $\sup_n \mathcal{J}(\mathcal{F}_n)$ is finite. Since $\sharp\mathcal{S}_n^{(2)} \leq \sqrt{n}\gamma_n(0,n)$, the above expression tends to zero as n goes to infinity. Therefore,

$$\left(\sum_{i\in\mathcal{S}_n^{(2)}} \phi_i(j,k)\right)_{(j,k)\in\mathcal{T}_n} = \left(\frac{2}{\sqrt{n}\gamma_n(0,n)}\sum_{i\in\mathcal{S}_n^{(2)}} I_{(j,k]}(i)\epsilon_i\theta_{in}\right)_{(j,k)\in\mathcal{T}_n} + o_p(1).$$

Concerning the first part $\sum_{i\in\mathcal{S}_n^{(1)}} \phi_i(.)$, note that

$$\mathbb{E}\Big(\sum_{i\in\mathcal{S}_n^{(1)}} \|\phi_i\|_{\mathcal{T}_n}^2\Big) = \mathbb{E}\Big(\sum_{i\in\mathcal{S}_n^{(1)}} (\phi_i(0,n))^2\Big) = \frac{1}{n\gamma_n(0,n)^2}\sum_{i\in\mathcal{S}_n^{(1)}} \mathbb{E}\big\{4\epsilon_i^2\theta_{in}^2 + (\epsilon_i^2-\sigma^2)^2\big\} = O(1)$$

while for any $u > 0$, the (uniform) Lindeberg condition

$$\mathbb{E}\Big(\sum_{i\in\mathcal{S}_n^{(1)}} I\big\{\|\phi_i\|_{\mathcal{T}_n}^2 > u\big\}\|\phi_i\|_{\mathcal{T}_n}^2\Big) = \quad \mathbb{E}\Big(\sum_{i\in\mathcal{S}_n^{(1)}} I\big\{\phi_i(0,n)^2 > u\big\}\phi_i(0,n)^2\Big) = o(1)$$

is easily seen to be satisfied. Since by construction the covariance function of $\tilde{D}_n/\gamma_n(0,n)$ is absolutely bounded by 1,

$$\sup_{\{\sharp\mathcal{T}_n^l = l | \mathcal{T}_n^l \subset \mathcal{T}_n\}} d_w\Big\{\mathcal{L}\Big(\sum_{k=1}^n \phi_k(t)\Big)_{t\in\mathcal{T}_n^l},\, \mathcal{N}\Big(0, \mathrm{cov}\Big(\sum_{k=1}^n \phi_k(t)\Big)_{t\in\mathcal{T}_n^l}\Big)\Big\} \longrightarrow 0$$

for all natural numbers l, due to the multivariate Lindeberg central limit theorem and the compactness of $[-1,1]$, which shows that condition (i) of Theorem 8 is satisfied. Condition (ii) results from the first part of the subsequent proof of Theorem 2. □

PROOF OF THEOREM 2 With the same argument as in the proof of Proposition 1, it is sufficient to prove the result with $\tilde{D}_n$ in place of D_n. Without loss of generality, we may further assume that $\mathrm{Var}\big(\tilde{D}_{2,n}(0,n)\big) = \gamma_n(0,n)^2 = 1$ by a simple rescaling argument. We begin with the situation where σ^2 is known ($\beta_n = 0$), i.e. we only consider the process $\tilde{D}_{2,n}$. By expanding the square, $\tilde{D}_{2,n}$ is of the general form

$$\sum_{i=1}^{n} \lambda_i (Z_i + \delta_i)^2 - \sum_{i=1}^{n} \lambda_i (1 + \delta_i^2)$$

with

$$\lambda_i = \lambda_{in}(j,k) = \frac{1}{\sqrt{n}} I_{(j,k]}(i) \quad \Big(|\lambda_i| \leq \frac{1}{\sqrt{n}}\Big), \qquad \delta_i = \delta_{in} = \frac{\theta_{in}}{\sigma}$$

and $Z_1, ..., Z_n$ i.i.d. $\mathcal{N}(0,1)$. Its expectation and variance are given by zero and

$$\mathrm{Var} \sum_{i=1}^{n} \lambda_i (Z_i + \delta_i)^2 \;=\; \sum_{i=1}^{n} 2\lambda_i^2 (1 + 2\delta_i^2),$$

respectively. Let the metric ρ_n on $\mathcal{T}_n \times \mathcal{T}_n$ be defined by

$$\rho_n\big((j,k),(j',k')\big)^2 \;:=\; \frac{1}{n} \sum_{i \in (j,k] \triangle (j',k']} \Big(\frac{4\,\theta_{in}^2}{\sigma^2} + 2\Big).$$

We first establish the following bound for the capacity numbers

$$D\Big(u\xi,\ \big\{(j,k) \in \mathcal{T}_n | \gamma_n(j,k) \leq \xi\big\},\ \rho_n\Big) \;\leq\; Au^{-4}\xi^{-2}$$

for some positive constant $A > 0$, independent of n, θ_n and ξ. For $0 \leq s \leq t \leq 1$ let

$$\mu_n([s,t]) \;:=\; \int_0^1 2I_{[s,t]}(x)(1 + 2\delta_n(x)^2) d\lambda(x),$$

where $\delta_n(.) := \sum_{i=1}^{n} \delta_{in} \cdot I_{[\frac{i-1}{n}, \frac{i}{n})}(.)$. Note that $\mu_n([0,1]) = 1$ in particular (by assumption). Then

$$\rho_n\big((j,k),(j',k')\big)^2 \;=\; \int_0^1 I_{[j/n,k/n] \Delta [j'/n,k'/n]} d\mu_n$$

and

$$\big\{(j,k) \in \mathcal{T}_n | \gamma_n(j,k) \leq \xi\big\} = \Big\{(j,k) \in \mathcal{T}_n \;\Big|\; \int_{j/n}^{k/n} d\mu_n \leq \xi^2\Big\}$$

for any $\xi \in (0,1]$. Let $\mathcal{S}_n = \{t_1, ..., t_m\} \subset [0,1]$ be a maximal subset with $t_1 = 0$ such that

$$\int_{t_i}^{t_{i+1}} d\mu_n(x) \;=\; \frac{u^2\xi^2}{2}.$$

Then $m \le 3/(u^2\xi^2)$. If now $(j,k),(j',k') \in \mathcal{T}_n$ with $j/n, j'/n \in [t_{i-1}, t_i]$ and $k/n, k'/n \in [t_{l-1}, t_l]$, $1 < i \le l \le m+1$ with $t_{m+1} = 1$ (if not already contained in $\mathcal{S}_n$), then

$$\rho_n\big((j,k),(j',k')\big) = \left(\int_0^1 \big(I_{[j/n,k/n]} - I_{[j'/n,k'/n]}\big)^2 d\mu_n\right)^{1/2} \le u\xi.$$

But $\xi^2 \ge \gamma_2(j,k)^2$ implies

$$\xi^2 \ge \int_{j/n}^{k/n} d\mu_n \ge \big(l-i-1\big)\frac{u^2\xi^2}{2}$$

which gives $l - i - 1 \le 2u^{-2}$. Hence,

$$\begin{aligned} D\Big(u\xi,\ \big\{(j,k) \in \mathcal{T}_n | \gamma_2(j,k) \le \xi\big\},\ \rho_n\Big) &\le \sharp\Big\{i < l \in \{1,...,m+1\}, l-i\ \le\ 1 + \frac{2}{u^2}\Big\} \\ &\le\ (m+1)\big(2+2u^{-2}\big)\ \le\ Au^{-4}\xi^{-2}. \end{aligned}$$

with $A > 0$ independent of n, θ_n and ξ.

The second exponential inequality in Proposition 6 gives

$$\mathbb{P}\Big(|\tilde{D}_{2,n}(j,k) - \tilde{D}_{2,n}(j',k')| \ge \rho_n\big((j,k),(j',k')\big)(4\eta + 1/2)\Big) \le 2\exp(-\eta),$$

which implies that

$$\mathbb{P}\Big(|\tilde{D}_{2,n}(j,k) - \tilde{D}_{2,n}(j',k')| \ge \rho_n\big((j,k),(j',k')\big)q\eta\Big)\ \le\ 2\exp(-\eta)$$

with $q = 4 + (2\log 2)^{-1}$. According to Theorem 7 and the subsequent remark 3 in Dümbgen and Walther (2008), there exists a constant $Q > 0$ such that

$$\lim_{\delta \searrow 0} \sup_{n \in \mathbb{N}} \mathbb{P}\left(\sup_{\rho_n\left((j,k),(j',k')\right) \le \delta} \frac{\big|\ \tilde{D}_{2,n}(j,k) - \tilde{D}_{2,n}(j',k')\ \big|}{\rho_n\big((j,k),(j',k')\big)\log\big(e/\rho_n\big((j,k),(j',k')\big)\big)} > Q \right)\ =\ 0,$$

implying in particular the stochastic equicontinuity condition (ii) of Theorem 8 in the appendix which has been left open in the proof of Proposition 1. Note that the same holds true with $\tilde{D}_{2,n}$ replaced by the approximating Gaussian process.

For notational convenience, let

$$T_n(\delta,\delta') := \sup_{\substack{(j,k)\in\mathcal{T}_n:\\ \delta < \gamma_n(j,k) \le \delta'}} \left\{ \frac{|\tilde{D}_{2,n}(j,k)|}{\gamma_n(j,k)} - C_{jkn} \right\}$$

for any $0 \le \delta < \delta' \le 1$ and analogously

$$S_n(\delta,\delta') := \sup_{\substack{(j,k)\in\mathcal{T}_n:\\ \delta < \gamma_n(j,k) \le \delta'}} \left\{ \frac{\big|\ W\big(\gamma_n(0,k)^2\big) - W\big(\gamma_n(0,j)^2\big)\ \big|}{\gamma_n(j,k)} - \Gamma_{jkn} \right\}$$

with $W(.)$ some Brownian motion on the unit interval. For any $\delta \in (0,1)$, $\sup_{\gamma_n(j,k)\geq\delta} |C_{jkn} - \Gamma_{jkn}| \to 0$ as n goes to infinity. Consequently by the proof of Proposition 1,

$$d_w\Big(\mathcal{L}\big(T_n(\delta,1)\big), \mathcal{L}\big(\big(S_n(\delta,1)\big)\Big) \;\longrightarrow\; 0 \quad (n\to\infty) \tag{17}$$

for any fixed $\delta \in (0,1)$. Note at this point that for the weak approximation by the dual bounded Lipschitz metric as defined in the appendix the continuous mapping theorem is not applicable in general. The statement follows since the mapping is Lipschitz continuous as long as $\delta > 0$.

Let

$$G_n(\eta,\delta) \;:=\; \frac{2\eta}{\sqrt{n}\delta} + \Big(\frac{4\eta^2}{n\delta^2} + 2\eta\Big)^{1/2}.$$

The Bernstein-type exponential inequality implies

$$\mathbb{P}\Big(|\tilde{D}_{2,n}(j,k)| \geq \gamma_n(j,k) G_n(\eta,\delta)\Big) \leq 2\exp(-\eta)$$

if $\gamma_n(j,k) \geq \delta$ for any fixed $\delta > 0$. The same holds true for the approximating Gaussian process with G_n replaced by $(2\eta)^{1/2}$. Then Theorem 8 in Dümbgen and Walther (2008) and its subsequent Remark imply

$$\lim_{\delta\searrow 0} \sup_n \mathbb{P}\big(T_n(0,\delta) \geq \epsilon\big) \;=\; 0 \quad \text{and} \quad \lim_{\delta\searrow 0} \sup_n \mathbb{P}\big(S_n(0,\delta) \geq \epsilon\big) \;=\; 0 \tag{18}$$

for $\varepsilon > 0$. For note that the variances $\gamma_n(j,k)^2/\gamma_n(0,n)^2$ appearing in the logarithms of the additive correction terms C_{jkn} can be replaced by $\max\big(\gamma_n(j,k)^2/\gamma_n(0,n)^2,\, 1/n\big)$ since the local covering numbers are bounded by n^2 anyway. Evidently,

$$\lim_{\delta\searrow 0} \sup_n \mathbb{P}\big(S_n(\delta,1) \leq -\epsilon\big) \;=\; 0 \tag{19}$$

for any fixed $\epsilon > 0$. Combining (17) – (19) yields

$$d_w\big(\mathcal{L}(T_n(0,1)), \mathcal{L}(S_n(0,1))\big) \;\to\; 0$$

as n goes to infinity.

So far we only considered the process $\tilde{D}_{2,n}$. If an additional estimation of σ^2 is involved, the process $\tilde{D}_{1,n}$ has to be taken into account as well. As the covariance function demonstrates, the standardized version of $\tilde{D}_{1,n}$ is stochastically bounded,

$$\sup_{(j,k)\in\mathcal{T}_n} \frac{|\tilde{D}_{1,n}(j,k)|}{\gamma_n(j,k)} = O_p(1).$$

Note that $\gamma_n^2(\cdot,\cdot)\cdot 2\beta_n^2 \geq \mathrm{Var}\big(\tilde{D}_{1,n}(\cdot,\cdot)\big)$. Let T_n' be defined as T_n above with $\tilde{D}_n$ in place of $\tilde{D}_{2,n}$, analogously define

$$S_n'(\delta,\delta') := \sup_{\substack{(j,k)\in\mathcal{T}_n:\\ \delta<\gamma_n(j,k)\leq\delta'}} \left(\frac{\big|\, W\big(\gamma_n(0,k)^2\big) - W\big(\gamma_n(0,j)^2\big) + 2\beta_n(k-j)/nZ \,\big|}{\gamma_n(j,k)} \;-\; \Gamma_{jkn} \right)$$

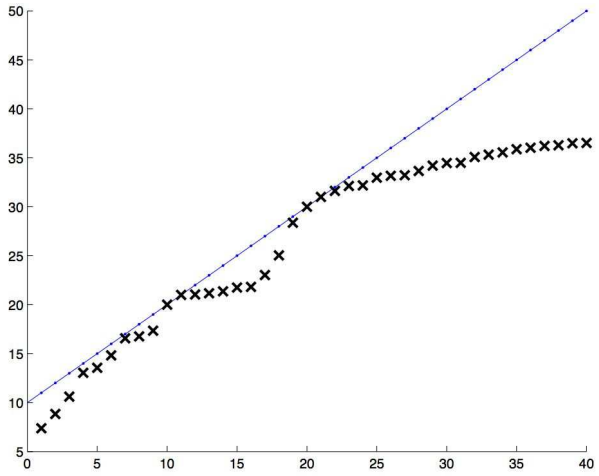

Figure 1: Construction of the coupling

for any $0 \leq \delta < \delta' \leq 1$. Claim (17) remains valid with T'_n and S'_n in place of T_n and S_n with the same argument. Furthermore,

$$\begin{aligned}
&\lim_{\delta \searrow 0} \limsup_{n\to\infty} \mathbb{P}(T'_n(0,\delta) \geq \epsilon) \\
&\quad \leq \lim_{\delta \searrow 0} \limsup_{n\to\infty} \mathbb{P}(T_n(0,\delta) \geq \epsilon/2) + \lim_{\delta \searrow 0} \limsup_{n\to\infty} \mathbb{P}\Bigg(\sup_{\substack{(j,k)\in\mathcal{T}_n: \\ 0<\gamma_n(j,k)\leq\delta}} \frac{|\tilde{D}_{1,n}(j,k)|}{\gamma_n(j,k)} \geq \epsilon/2 \Bigg) \\
&\quad = 0,
\end{aligned}$$

using $\lim_{\delta\searrow 0} \sup_{0<\gamma_n(j,k)<\delta} \mathrm{Var}\big(\tilde{D}_{1,n}(j,k)\big)/\gamma_n(j,k) = 0$. Analogously, the conclusion is true with S'_n in place of T'_n. Clearly, (19) follows for S'_n as well, which completes the proof. □

PROOF OF PROPOSITION 3 The main ingredient is a well-known representation of non-central χ^2 distributions as Poisson mixtures of central χ^2 distributions. Precisely,

$$\chi^2_k(\delta^2) = \sum_{j=0}^{\infty} e^{-\delta^2/2} \frac{(\delta^2/2)^j}{j!} \cdot \chi^2_{k+2j},$$

as can be proved via Laplace transforms. Now we define 'time points'

$$t_{kn} := \sum_{i=1}^{k} \theta_{in}^2/\sigma^2 \quad \text{and} \quad t^*_{kn} := t_{j(n)n} + k - j(n)$$

with $j(n)$ any fixed index in $\mathcal{K}_n(\theta_n)$. This construction entails that $t^*_{kn} \geq t_{kn}$ with equality if, and only if, $k \in \mathcal{K}_n(\theta_n)$.

Figure 1 illustrates this construction. It shows the time points t_{kn} (crosses) and t^*_{kn} (dots and line) versus k for a hypothetical signal $\theta_n \in \mathbb{R}^{40}$ with $\sigma = 1$. Note that in this example, $\mathcal{K}_n(\theta_n)$ is given by $\{10, 11, 20, 21\}$.

Let Π, G_1, G_2, ..., G_n, Z_1, Z_2, Z_3, ...and S^2 be stochastically independent random variables, where $\Pi = (\Pi(t))_{t\geq 0}$ is a standard Poisson process, G_i and Z_j are standard Gaussian random variables, and $S^2 \sim \chi^2_m$. Then one can easily verify that

$$\begin{aligned}\tilde{T}_{jkn} &:= \frac{m}{(k-j)S^2}\Big(\sum_{i=j+1}^{k} G_i^2 + \sum_{s=2\Pi(t_{jn}/2)+1}^{2\Pi(t_{kn}/2)} Z_s^2\Big),\\ \tilde{T}^*_{jkn} &:= \frac{m}{(k-j)S^2}\Big(\sum_{i=j+1}^{k} G_i^2 + \sum_{s=2\Pi(t^*_{jn}/2)+1}^{2\Pi(t^*_{kn}/2)} Z_s^2\Big)\end{aligned}$$

define random variables $(\tilde{T}_{jkn})_{0\leq j<k\leq n}$ and $(\tilde{T}^*_{jkn})_{0\leq j<k\leq n}$ with the desired properties. □

PROOF OF THEOREM 4 Recall that

$$\gamma_n(j,k)^2 \;=\; \frac{1}{n}\sum_{i=j+1}^{k}\big(4\theta_{in}^2/\sigma^2 + 2\big), \tag{20}$$

which equals $\gamma^*_n(j,k)^2 := 6|k-j|/n$ in case of $\theta_n = \theta^*_n$. Without loss of generality let $\sigma = 1$. If $\hat{\sigma}_n^2$ satisfies condition (A), Proposition 3 yields that

$$\mathbb{P}_{\theta_n}\Big(\mathcal{K}_n(\theta_n) \subset \hat{\mathcal{K}}_{n,\alpha}\Big) \;\geq\; 1-\alpha.$$

The statements about the asymptotic behavior of $\kappa_{n,\alpha}$ are an immediate consequence of Theorem 2. Our next goal is to establish the oracle inequality (6), where the stochastic order terms o_p and O_p are supposed to be independent of $(\theta_n)_{n\in\mathbb{N}}$. First note that

$$\frac{1}{\sqrt{n}}\gamma_n(j,k)C_{jkn} \;\leq\; K\,\frac{\log n}{n} \;+\; \frac{\gamma_n(j,k)}{\sqrt{n}}\,\Gamma(1/n). \tag{21}$$

Here and in what follows, K denotes some universal constant, independent of $(\theta_n)_{n\in\mathbb{N}}$, j, k and n. Its value may be different in different expressions. By the definition of $\hat{\mathcal{K}}_{n,\alpha}$,

$$\hat{R}(k) \;\leq\; \hat{R}(j) + \frac{\hat{\sigma}_n^2\gamma^*_n(j,k)}{\sqrt{n}}\Big(\Gamma\big(|k-j|/n\big) + \kappa_{n,\alpha}\Big) + \frac{5\,\hat{\sigma}_n^2}{n}\Gamma\big(|k-j|/n\big)^2 \tag{22}$$

for all $k \in \hat{\mathcal{K}}_{n,\alpha}$ and $j \neq k$, in particular for every $j \in \mathcal{K}_n(\theta_n)$. As a consequence of the tightness shown in Theorem 2,

$$\begin{aligned}R(k) - R(j) \;&\leq\; \hat{R}(k) - \hat{R}(j) \;+\; \hat{\sigma}_n^2 K\,\frac{\log n}{n} \;+\; \hat{\sigma}_n^2\frac{\gamma_n(j,k)}{\sqrt{n}}\Big(\Gamma(1/n) + Z'_n\Big)\\ &=\; \hat{R}(k) - \hat{R}(j) \;+\; \big(1 + O_p(n^{-1/2})\big)\Big\{K\,\frac{\log n}{n} \;+\; \frac{\gamma_n(j,k)}{\sqrt{n}}\Big(\Gamma(1/n) + Z'_n\Big)\Big\}\end{aligned} \tag{23}$$

for some random variable $Z'_n = O_p(1)$, independent of j, k, θ_n. Thus (21 – 23) imply for any $j \in \mathcal{K}_n(\theta_n)$ and $k \in \hat{\mathcal{K}}_{n,\alpha}$,

$$R(k) - R(j) \;\le\; (K + Z_n)\frac{\log n}{n} \;+\; \frac{\Gamma(1/n)}{\sqrt{n}}\,\big(1 + Z_n\big)\Big(\gamma_n(j,k) + \gamma_n^*(j,k)\Big) \qquad (24)$$

for some positive constant K and a positive random variable $Z_n = o_p(1)$, independent of j, k and θ_n. Using that $\sqrt{x} + \sqrt{y} \le \sqrt{x/\lambda + y/(1-\lambda)}$ for any $x, y \ge 0$ and $\lambda \in (0,1)$,

$$\begin{aligned}\sqrt{n}\Big(\gamma_n(j,k) + \gamma_n^*(j,k)\Big) \;&\le\; \sqrt{10\frac{1}{n}\sum_{i=\min(j,k)+1}^{\max(j,k)} \theta_{in}^2 \;+\; 15\frac{|k-j|}{n}} \\ &\le\; \sqrt{15(R_n(j) + R_n(k))}\end{aligned}$$

setting $\lambda = 2/5$. This is easily shown to entail that

$$R(k) \;\le\; R(j) \;+\; \frac{\Gamma(1/n)}{n}\big(1 + Z_n\big)\sqrt{30}\sqrt{R(j)} \;+\; (K + Z_n)^2\frac{\log n}{n}.$$

Let $L(j,k) := L(k) - L(j)$ and $R(j,k) := R(k) - R(j)$. First note that for any $j < k$,

$$\big(L(j,k) - R(j,k)\big) \;=\; \frac{1}{n}\sum_{i=j+1}^{k}(\varepsilon_i^2 - 1).$$

Analogously to the proof of Theorem 2, there exist a sequence of random variables (Z_n) and some constant K, both independent of j, n and (θ_n) with $Z_n = O_p(1)$ such that

$$\big|\, L(j,k) - R(j,k)\,\big| \;\le\; \left\{K\,\frac{\log n}{n} \;+\; \frac{\gamma_n^+(j,k)}{\sqrt{n}}\Big(\Gamma(1/n) + Z_n\Big)\right\}$$

with $\gamma_n^+(j,k)^2 := 2|k-j|/n$. Consequently, for any $j \in \hat{\mathcal{K}}_{n,\alpha}$ and $j \ne k$,

$$\begin{aligned}L(j) - L(k) \;&=\; (L - R)(j,k) + R(j,k) \\ &\le\; (K + o_p(1))\frac{\log n}{n} \;+\; \frac{\Gamma(1/n)}{\sqrt{n}}\,\big(1 + o_p(1)\big)\Big(\gamma_n^+(j,k) + \gamma_n^*(j,k) + \gamma_n(j,k)\Big).\end{aligned}$$

But

$$\sqrt{n}\Big(\gamma_n^+(j,k) + \gamma_n^*(j,k) + \gamma_n(j,k)\Big) \;\le\; \sqrt{K(R(j) + R(k))},$$

while the above inequality applied for $|L(j) - R(j)|$ shows that

$$R(j) - \left\{K\,\frac{\log n}{n} \;+\; \frac{\sqrt{2R(j)}}{\sqrt{n}}\Big(\Gamma(1/n) + Z_n\Big)\right\} \;\le\; L(j),$$

whence $\tilde{R}(j) \le 2\tilde{L}(j) + (K + Z_n)^2(\log n)/n$. Therefore,

$$L(j) \;\le\; L(k) \;+\; \frac{\Gamma(1/n)}{\sqrt{n}}\big(1 + o_p(1)\big)\sqrt{KL(k)} \;+\; (K + o_p(1))\frac{\log n}{n}. \qquad \square$$

PROOF OF THEOREM 5 Let $\mu_n := \log M_n$. The application of inequality (15) in Corollary 7 to the tripel $(\#J, T_n(J), \alpha/M_n)$ in place of (n, t, α) yields bounds for $\hat{\delta}^2_{n,\alpha,l}(J)$ and $\hat{\delta}^2_{n,\alpha,u}(J)$ in terms of $\hat{\delta}^2_n(J) := (T_n(J) - \#J)_+$. Then we apply (13–14) to $T_n(J)$, where (n, δ^2, α) is to be replaced with $(\#J, \delta^2_n(J), \alpha'/M_n)$ for any fixed $\alpha' \in (0, 1)$. Using the fact that for arbitrary constants $a, b, c > 0$, the function $h(x) := x + \sqrt{a + bx} + c$, $x \geq 0$, satisfies the inequality

$$h(h(x)) \;\leq\; x + 2\sqrt{a + bx} + \big(2c + b/2 + \sqrt{bc}\big),$$

we obtain finally

$$\left.\begin{array}{c} \hat{\delta}^2_{n,\alpha,u}(J) - \delta^2_n(J) \\ \delta^2_n(J) - \hat{\delta}^2_{n,\alpha,l}(J) \end{array}\right\} \;\leq\; (1 + o_p(1))\sqrt{(16\#J + 32\,\delta^2_n(J))\mu_n} + (K + o_p(1))\mu_n \quad (25)$$

for all $J \in \mathcal{M}_n$. Here, K denotes denotes some universal constant, independent of σ, $(\theta_n)_{n\in\mathbb{N}}$, C, D and n. Its value may be different in different expressions. We consider $\tilde{R}_n(C) := (n/\sigma^2)R_n(C) = \delta^2_n(C^{\mathrm{c}}) + \#C$. It follows from the definition of the confidence region $\hat{\mathcal{K}}_{n,\alpha}$ that for arbitrary $C \in \hat{\mathcal{K}}_{n,\alpha}$ and $D \in \mathcal{C}_n$,

$$\begin{aligned} \tilde{R}_n(C) - \tilde{R}_n(D) &= \delta^2_n(D \setminus C) - \delta^2_n(C \setminus D) + \#C - \#D \\ &= (\delta^2_n - \hat{\delta}^2_{n,\alpha,l})(D \setminus C) + (\hat{\delta}^2_{n,\alpha,u} - \delta^2_n)(C \setminus D) \\ &\quad - \big(\hat{\delta}^2_{n,\alpha,u}(C \setminus D) - \hat{\delta}^2_{n,\alpha,l}(D \setminus C) + \#D - \#C\big) \\ &\leq (\delta^2_n - \hat{\delta}^2_{n,\alpha,l})(D \setminus C) + (\hat{\delta}^2_{n,\alpha,u} - \delta^2_n)(C \setminus D). \end{aligned}$$

Moreover, according to (25) the latter bound is not larger than

$$\begin{aligned} &(1 + o_p(1))\Big\{\sqrt{\big(16\#(D \setminus C) + 32\delta^2_n(D \setminus C)\big)\mu_n} + \sqrt{\big(16\#(C \setminus D) + 32\delta^2_n(C \setminus D)\big)\mu_n}\Big\} \\ &\qquad + (K + o_p(1))\mu_n \\ &\leq\; (1 + o_p(1))\sqrt{2\mu_n\big(16\#D + 32\delta^2_n(C^{\mathrm{c}}) + 16\#C + 32\delta^2_n(D^{\mathrm{c}})\big)} \;+\; (K + o_p(1))\mu_n \\ &\leq\; 8\sqrt{\mu_n\big(\tilde{R}_n(C) + \tilde{R}_n(D)\big)}(1 + o_p(1)) \;+\; (K + o_p(1))\mu_n. \end{aligned}$$

Thus we obtain the quadratic inequality

$$\tilde{R}_n(C) - \tilde{R}_n(D) \;\leq\; 8\sqrt{\mu_n\big(\tilde{R}_n(C) + \tilde{R}_n(D)\big)}(1 + o_p(1)) \;+\; (K + o_p(1))\mu_n,$$

which is easily shown to entail that

$$\tilde{R}_n(C) \;\leq\; \tilde{R}_n(D) + 8\sqrt{2}\sqrt{\tilde{R}_n(D)\mu_n}(1 + o_p(1)) \;+\; (K + o_p(1))^2\mu_n.$$

This yields the assertion about the risks.

As for the losses, note that $\tilde{L}_n(\cdot) := (n/\sigma^2)L_n(\cdot)$ and $\tilde{R}_n(\cdot)$ are closely related in that

$$(\tilde{L}_n - \tilde{R}_n)(D) \;=\; \sum_{i\in D} \epsilon^2_{in}/\sigma^2 - \#J$$

for arbitrary $D \in \mathcal{C}_n$. Hence we may utilize (13–14) with $(\#D, 0, \alpha'/\mu_n)$ in place of (n, δ^2, α) to complement (25) with the following observation:

$$-A\sqrt{\#D\mu_n} \;\le\; \tilde{L}_n(D) - \tilde{R}_n(D) \;\le\; A\sqrt{\#D\mu_n} + A\mu_n \quad \text{for all } D \in \mathcal{C}_n \tag{26}$$

with probability tending to one as $n \to \infty$ and $A \to \infty$. Note also that (26) implies the inequality $\tilde{R}_n(D) - A\sqrt{\tilde{R}_n(D)\mu_n} \le \tilde{L}_n(D)$, whence

$$\tilde{R}_n(D) \;\le\; 2\tilde{L}_n(D) + A^2\mu_n/2 \quad \text{for all } D \in \mathcal{C}_n$$

Assuming that both (25) and (26) hold for some large but fixed A, we may conclude that for arbitrary $C \in \hat{\mathcal{K}}_{n,\alpha}$ and $D \in \mathcal{C}_n$,

$$\begin{aligned}
&\tilde{L}_n(C) - \tilde{L}_n(D) \\
&\quad= \;(\tilde{L}_n - \tilde{R}_n)(C) - (\tilde{L}_n - \tilde{R}_n)(D) + \tilde{R}_n(C) - \tilde{R}_n(D) \\
&\quad\le\; A\sqrt{2(\#C + \#D)\mu_n} + A\sqrt{2\mu_n\big(\tilde{R}_n(C) + \tilde{R}_n(D)\big)} + 4A\mu_n \\
&\quad\le\; 2A\sqrt{2\mu_n\big(\tilde{R}_n(C) + \tilde{R}_n(D)\big)} + 4A\mu_n \\
&\quad\le\; A'\sqrt{2\mu_n\big(\tilde{L}_n(C) + \tilde{L}_n(D)\big)} + 2A'\mu_n
\end{aligned}$$

for some constant $A' = A'(A)$. Again this inequality entails that

$$\tilde{L}_n(C) \;\le\; \tilde{L}_n(D) + A'\sqrt{2\tilde{L}_n(D)\mu_n} + 4A'^2\mu_n. \qquad \square$$

6 Auxiliary results

This section collects the main auxiliary results in the context of empirical processes which are useful to establish our results. They are formulated in quite an abstract framework to avoid notational expenditure.

For any pseudo–metric space $(\mathcal{X}, d)$ and $u > 0$, we define the covering number

$$N(u, \mathcal{X}, d) := \min\big\{\sharp\mathcal{X}_o \,\big|\, \mathcal{X}_o \subset \mathcal{X},\, \inf_{x_o \in \mathcal{X}_o} d(x, x_o) \le u \text{ for all } x \in \mathcal{X}\big\}.$$

The proof of Proposition 1 requires the following definition of uniform covering numbers. For some set $\mathcal{T}$, let $\mathcal{F} \subset [0,1]^{\mathcal{T}}$. For any discrete probability measure P on $\mathcal{T}$, consider the pseudo-distance $d_P(f,g)^2 := \int (f-g)^2\, dP$ for $f, g \in \mathcal{F}$. Then the uniform covering numbers of $\mathcal{F}$ are defined as

$$\mathcal{N}(u, \mathcal{F}) := \sup_P N(u, \mathcal{F}, d_P)$$

for $u > 0$, where the supremum is running over all discrete probability measures P on $\mathcal{T}$. If in particular $\mathcal{T} = \mathcal{C}_n$ and $\mathcal{F} = \mathcal{F}_n = \big\{I_{[0,t]} | t \in \mathcal{C}_n\big\}$, then elementary calculations show that $N(u, \mathcal{F}_n) \le 1 + u^{-2} \le 2u^{-2}$ for $0 < u \le 1$.

It is well-known that convergence in distribution of random variables with values in a separable metric space may be metrized by the dual bounded Lipschitz distance. Now we adapt the latter distance for stochastic processes. Let $\ell_\infty(\mathcal{T})$ be the space of bounded functions $x : \mathcal{T} \to \mathbb{R}$, equipped with supremum norm $\|\cdot\|_\infty$. For two stochastic processes X and Y on $\mathcal{T}$ with bounded sample paths we define

$$d_w(X, Y) \ := \ \sup_{f \in \mathcal{H}(\mathcal{T})} \big|\mathbb{E}^* f(X) - \mathbb{E}^* f(Y)\big|,$$

where $\mathbb{P}^*$ and $\mathbb{E}^*$ denote outer probabilities and expectations, while $\mathcal{H}(\mathcal{T})$ is the family of all funtionals $f : \ell_\infty(\mathcal{T}) \to \mathbb{R}$ such that

$$|f(x)| \ \le \ 1 \quad \text{and} \quad |f(x) - f(y)| \ \le \ \|x - y\|_\infty \quad \text{for all } x, y \in \ell_\infty(\mathcal{T}).$$

If d is a pseudo-metric on $\mathcal{T}$, then the modulus of continuity $w(x, \delta|d)$ of a function $x \in l_\infty(\mathcal{T})$ is defined as

$$w(x, \delta|d) := \sup_{s,t \in \mathcal{T} : d(s,t) \le \delta} |x(s) - x(t)|.$$

Furthermore, $\mathcal{C}_u(\mathcal{T}, d)$ denotes the set of uniformly continuous functions on $(\mathcal{T}, d)$, that is

$$\mathcal{C}_u(\mathcal{T}, d) \ = \ \Big\{ x \in l_\infty(\mathcal{T}) : \lim_{\delta \searrow 0} w(x, \delta|d) = 0 \Big\}.$$

Theorem 8. *For $n = 1, 2, 3, \ldots$ let $X_n = \big(X_n(t)\big)_{t \in \mathcal{T}_n}$ and $Y_n = \big(Y_n(t)\big)_{t \in \mathcal{T}_n}$ be stochastic processes on a metric space $(\mathcal{T}_n, \rho_n)$ with bounded sample paths. Then*

$$d_w(X_n, Y_n) \ \to \ 0$$

provided that the following three conditions are satisfied:

(i) For any integer $k > 0$,

$$\sup_{A_n \subset \mathcal{T}_n : \sharp A_n \le k} d_w\Big(X_n\big|_{A_n}, Y_n\big|_{A_n}\Big) \ \longrightarrow \ 0;$$

(ii) for each positive number ϵ,

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} \mathbb{P}^*\big(w(Z_n, \delta|\rho_n) > \epsilon\big) \ = \ 0 \quad \text{for } Z_n = X_n, Y_n;$$

(iii) for all $u > 0$, $\quad \sup_n N(u, \rho_n, \mathcal{T}_n) < \infty$.

PROOF For every natural number k let $\mathcal{T}_n^k$ be some maximal subset of $\mathcal{T}_n$ such that $\rho_n(t, t') \ge 1/k$ for any $t, t' \in \mathcal{T}_n^k$, and $\mathcal{T}_n^1 \subset \mathcal{T}_n^2 \subset \mathcal{T}_n^3 \subset \cdots$. Consequently, $\rho(t, \mathcal{T}_n^k) \le 1/k$ for every $t \in \mathcal{T}_n$. Now define

$$\lambda_n^k(t, u) := \frac{\big(1 - k\rho_n(t, u)\big)^+}{\sum_{v \in \mathcal{T}_n^k} \big(1 - k\rho_n(t, v)\big)^+}$$

for all $t \in \mathcal{T}_n$ and $u \in \mathcal{T}_n^k$. Note that $0 \le \lambda_n^k(\cdot, u) \in \mathcal{C}_u(\mathcal{T}_n, \rho_n)$, $\sum_{u\in\mathcal{T}_n^k} \lambda_n^k(\cdot, u) \equiv 1$, and $\lambda_n^k(t, u) = 0$ if $\rho_n(t, u) \ge 1/k$. Now let

$$\pi_k^n : l_\infty(\mathcal{T}_n) \ \ (\text{or } \ l_\infty(\mathcal{T}_n^k)) \to \mathcal{C}_u(\mathcal{T}_n, \rho_n)$$

be defined by

$$\pi_k^n f := \sum_{u\in\mathcal{T}_n^k} f(u)\lambda_k^n(., u)$$

Then π_k^n is is some linear map such that

$$\|\pi_k^n f\|_{\sup} \le \|f\|_{\mathcal{T}_n^k} \ \text{ for all } \ \ f \in l_\infty(\mathcal{T}_n) \cup l_\infty(\mathcal{T}_n^k) \ \text{ and}$$

$$\|f - \pi_k^n f\|_{\sup} \le w(f, 1/k|\rho_n) \ \ \text{ for all } \ f \in l_\infty(\mathcal{T}_n).$$

Especially, π_k^n is Lipschitz continuous with constant 1, because for $f, g \in l_\infty(\mathcal{T}_n)$,

$$\big|\, \pi_k^n f - \pi_k^n g \,\big| \ = \ \Big|\, \sum_{u\in\mathcal{T}_n^k} \big(f(u) - g(u)\big)\lambda_k^n(., u) \,\Big| \ \le \sup_{u\in\mathcal{T}_n^k} \big|\, f(u) - g(u) \,\big| \sum_{u\in\mathcal{T}_n^k} \lambda_k^n(., u) \le \big\|\, f - g \,\big\|_{\mathcal{T}_n}.$$

Hence note that for all

$$f : l_\infty(\mathcal{T}_n) \text{ or } l_\infty(\mathcal{T}_n^k) \to [0, 1]$$

which are Lipschitz continuous with constant L, the composition $f \circ \pi_k^n$ again takes values in $[0, 1]$ and is Lipschitz continuous with constant L.

Then

$$\begin{aligned}
&\sup_{f\in\mathcal{H}(\mathcal{T}_n)} |\mathbb{E}^* f(X_n) - \mathbb{E}^* f(Y_n)| \\
&\qquad \le \sup_{f\in\mathcal{H}(\mathcal{T}_n)} \mathbb{E}^* |f(X_n) - f(\pi_k^n X_n)| + \sup_{f\in\mathcal{H}(\mathcal{T}_n)} \mathbb{E}^* |f(Y_n) - f(\pi_k^n Y_n)| \\
&\qquad\qquad + \sup_{f\in\mathcal{H}(\mathcal{T}_n)} |\mathbb{E}^* f(\pi_k^n X_n) - \mathbb{E}^* f(\pi_k^n Y_n)|
\end{aligned}$$

Because of assumption (i), $\sup_n \sharp\mathcal{T}_n^k < \infty$ by (iii) and $\{f \circ \pi_k^n | f \in \mathcal{H}(\mathcal{T}_n)\} \subset \mathcal{H}(\mathcal{T}_n^k)$,

$$\sup_{f\in\mathcal{H}(\mathcal{T}_n)} |\mathbb{E}^* f(\pi_k^n X_n) - \mathbb{E}^* f(\pi_k^n Y_n)| \underset{n\to\infty}{\longrightarrow} 0.$$

Let $\epsilon > 0$. Because of (ii), there exists some natural number $k = k(\epsilon)$ such that

$$\limsup_{n\to\infty} \mathbb{P}^*\big(w(Z_n, 1/k) > \epsilon\big) \le \epsilon \ \ \text{ for } Z_n = X_n, Y_n.$$

With this choice of k,

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{H}(\mathcal{T}_n)} \mathbb{E}^* |f(Z_n) - f(\pi_k^n Z_n)| \le \limsup_{n\to\infty} \Big(\epsilon + \mathbb{P}^*\big(w(Z_n, 1/k) \ge \epsilon\big)\Big) \le 2\epsilon.$$

This yields the desired result. □

Proposition 9. *Let* $\big(X_n(t)\big)_{t\in\mathcal{T}_n}$ *and* $\big(Y_n(t)\big)_{t\in\mathcal{T}_n}$ *independent stochastic processes on a metric space* $(\mathcal{T}_n,\rho_n)$*. Let* $\big(X'_n(t)\big)_{t\in\mathcal{T}_n}$ *and* $\big(Y'_n(t)\big)_{t\in\mathcal{T}_n}$ *be independent stochastic processes such that*

$$d_w\big(X_n,X'_n\big)\ \to\ 0\ \textit{and}\ d_w\big(Y_n,Y'_n\big)\ \to\ 0.$$

Assume that $\mathcal{T}_n$ *is either countable or all processes have continuou sample paths with respect to* ρ_n*. Then*

$$d_w\big(X_n+Y_n,X'_n+Y'_n\big)\ \to\ 0.$$

References

[1] BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* ***32***, 528–551.

[2] BERAN, R. (1996). Confidence sets centered at C_p estimators. *Ann. Inst. Statist. Math.* ***48***, 1-15.

[3] BERAN, R. (2000). REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Assoc.* ***95***, 155-169.

[4] BERAN, R. and DÜMBGEN, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* ***26***, 1826–1856.

[5] CAI, T.T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequailty approach. *Ann. Statist.* ***26***, 1783–1799.

[6] CAI, T.T. (2002). On block thresholding in wavelet regression: adaptivity, block size, and threshold level. *Statist. Sin.* ***12***, 1241–1273.

[7] CAI, T.T. and LOW, M.G. (2006). Adaptive confidence balls. *Ann. Statist.* ***34***, 202–228.

[8] CAI, T.T. and LOW, M.G. (2007). Adaptive estimation and confidence intervals for convex functions and monotone functions. *Manuscript in preparation.*

[9] DAHLHAUS, R. and POLONIK, W. (2006). Nonparametric quasi-maximum likelihood estimation for Gaussian locally stationary processes. *Ann. Statist.* ***34***, 2790–2824.

[10] DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* ***81***, 425–455.

[11] DONOHO, D.L. and JOHNSTONE, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *JASA* ***90***, 1200–1224.

[12] DONOHO, D.L. and JOHNSTONE, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* ***26***, 879–921.

[13] DÜMBGEN, L. (2002). Application of local rank tests to nonparametric regression. *J. Nonpar. Statist.* ***14***, 511–537.

[14] DÜMBGEN, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli* ***9***, 423–449.

[15] DÜMBGEN, L. and SPOKOINY, V.G. (2001). Multiscale testing of qualitative hypotheses. *Ann. Statist.* ***29***, 124–152.

[16] DÜMBGEN, L. and WALTHER, G. (2006, revised 2007). Multiscale inference about a density. Technical report 56, IMSV, University of Bern.

[17] EFROMOVICH, S. (1998). Simultaneous sharp estimation of functions and their derivatives. *Ann. Statist.* ***26***, 273–278.

[18] FUTSCHIK, A. (1999). Confidence regions for the set of global maximizers of nonparametrically estimated curves. *J. Statist. Plann. Inf.* ***82***, 237–250.

[19] GENOVESE, C.R. and WASSERMANN, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.* ***33***, 698–729.

[20] HENGARTNER, N.W. and STARK, P.B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* ***23***, 525–550.

[21] HOFFMANN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression (with discussion). *Ann. Statist.* ***30***, 325–396.

[22] LEPSKI, O.V., MAMMEN, E. and SPOKOINY, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* ***25***, 929–947.

[23] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* ***17***, 1001–1008.

[24] POLYAK, B.T. and TSYBAKOV, A.B. (1991). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Probab. Appl.* ***35***, 293–306.

[25] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* ***34***, 229–253.

[26] STONE, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* ***12***, 1285–1297.